Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models

Jimmy Olsson, Olivier Cappé, Randal Douc, Eric Moulines

Bernoulli **14**(1), 2008, 155–179
DOI: 10.3150/07-BEJ6150

Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models

JIMMY OLSSON[1,*], OLIVIER CAPPÉ[1,**], RANDAL DOUC[2] and ÉRIC MOULINES[1,†]

[1] *Ecole Nationale Supérieure des Télecommunications, France. E-mail:* [*] *olsson@enst.fr;* [**] *cappe@enst.fr;* [†] *moulines@enst.fr*
[2] *CMAP, Ecole Polytechnique, France. E-mail: douc@cmapx.polytechnique.fr*

This paper concerns the use of sequential Monte Carlo methods (SMC) for smoothing in general state space models. A well-known problem when applying the standard SMC technique in the smoothing mode is that the resampling mechanism introduces degeneracy of the approximation in the path space. However, when performing maximum likelihood estimation via the EM algorithm, all functionals involved are of additive form for a large subclass of models. To cope with the problem in this case, a modification of the standard method (based on a technique proposed by Kitagawa and Sato) is suggested. Our algorithm relies on forgetting properties of the filtering dynamics and the quality of the estimates produced is investigated, both theoretically and via simulations.

Keywords: EM algorithm; exponential family; particle filters; sequential Monte Carlo methods; state space models; stochastic volatility model

1. Introduction

In this paper, we study SMC methods for smoothing in nonlinear state space models. We consider a bivariate process (X,Y), where $X \triangleq \{X_k; k \geq 0\}$ is a homogeneous discrete-time Markov chain taking values in some state space $(\mathsf{X}, \mathcal{X})$. We let $(Q_\theta, \theta \in \Theta \subseteq \mathbb{R}^d)$ and ν denote the Markov transition kernel and the initial distribution of X, respectively. The family $\{Q_\theta(x,\cdot);\ x \in \mathsf{X},\ \theta \in \Theta\}$ is assumed to be dominated by the probability measure μ on $(\mathsf{X}, \mathcal{X})$ and we denote by $q_\theta(x,\cdot)$ the corresponding Radon–Nikodym derivatives. In this framework, X is not observed and measurements must be made through the process $Y \triangleq \{Y_k; k \geq 0\}$ taking values in some measurable space $(\mathsf{Y}, \mathcal{Y})$. These observed variables are conditionally independent, given the sequence $\{X_k; k \geq 0\}$, and the conditional distribution of Y_k depends only on X_k. We denote by $\mathcal{G}_k$ the σ-algebra generated by the

observed process from time zero to k. Furthermore, there exist, for all $x \in \mathsf{X}$ and $\theta \in \Theta$, a density function $y \mapsto g_\theta(x, y)$ and a measure λ on $(\mathsf{Y}, \mathcal{Y})$ such that, for $k \geq 0$,

$$\mathbb{P}_\theta(Y_k \in A | X_k = x) = \int_A g_\theta(x, y)\lambda(\mathrm{d}y) \qquad \text{for all } A \in \mathcal{Y}.$$

Here, we have written $\mathbb{P}_\theta$ for the law of the bivariate Markov chain $\{(X_k, Y_k); k \geq 0\}$ under the model parameterized by $\theta \in \Theta$ and we denote by $\mathbb{E}_\theta$ the associated expectation.

For $i \leq j$, let $X_{i:j} \triangleq (X_i, \ldots, X_j)$; similar vector notation will be used for other quantities. In many situations, it is required to compute expectation values of the form $\mathbb{E}_\theta[t_n(X_{0:n})|\mathcal{G}_n]$, where t_n is a real-valued, measurable function. In this paper, we focus on the case where t_n is an *additive functional* given by

$$t_n(x_{0:n}) = \sum_{k=0}^{n-1} s_k(x_{k:k+1}), \tag{1}$$

where $\{s_k; k \geq 0\}$ is a sequence of measurable functions (which may depend on the observed values $Y_{0:n}$).

As an example of when smoothing of such additive functionals is important, consider the case of maximum likelihood estimation via the *EM algorithm*. Having an initial estimate θ' of the parameter vector available, an improved estimate is obtained (we refer to Cappé *et al.* [2], Section 10.2.3) by means of computation and maximization of $\mathcal{Q}(\theta; \theta')$ with respect to θ, where $\mathcal{Q}(\theta; \theta')$ is defined by

$$\begin{aligned}\mathcal{Q}(\theta; \theta') \triangleq{}& \mathbb{E}_{\theta'}\left[\sum_{k=0}^{n-1} \log q_\theta(X_k, X_{k+1}) \Big| \mathcal{G}_n\right] + \mathbb{E}_{\theta'}\left[\sum_{k=0}^{n} \log g_\theta(X_k, Y_k) \Big| \mathcal{G}_n\right] \\ &+ \mathbb{E}_{\theta'}[\log \nu(X_0)|\mathcal{G}_n].\end{aligned}$$

This procedure is recursively repeated in order to obtain convergence to a stationary point $\theta_\star$ of the *log-likelihood function* $\ell_{\nu,n}(\theta; Y_{0:n}) \triangleq \log \mathrm{L}_{\nu,n}(\theta; Y_{0:n})$, where, for $y_{0:n} \in \mathsf{Y}^{n+1}$,

$$\mathrm{L}_{\nu,n}(\theta; y_{0:n}) \triangleq \int_{\mathsf{X}^{n+1}} g_\theta(x_0, y_0)\nu(x_0) \prod_{k=1}^{n} q_\theta(x_{k-1}, x_k) g_\theta(x_k, y_k) \mu^{\otimes(n+1)}(\mathrm{d}x_{0:n}).$$

The computation of smoothed sum functionals of the above form will also be the key issue when considering direct maximum likelihood estimation via the *score function* $\nabla_\theta \ell_{\nu,n}(\theta; y_{0:n})$; again see Cappé *et al.* ([2], Section 10.2.3) for details.

By applying Bayes' formula, it is straightforward to derive recursive formulas for expectations of the additive type discussed above. However, tractable closed form solutions are available only if the state space X is finite or the model is linear and Gaussian.

SMC methods (also known as *particle filtering methods*) constitute a class of algorithms that are well suited for providing approximate solutions of the smoothing and filtering recursions. In recent years, SMC methods have been applied, sometimes very

successfully, in many different fields (see Doucet *et al.* [6] and Ristic *et al.* [14] and the references therein). A well-known problem when applying SMC methods to sample the joint smoothing distribution is that the resampling mechanism of the particle filter introduces degeneracy of the particle trajectories. Doucet *et al.* [7] suggest a procedure where this is avoided through an additional resampling pass in the time-reversed direction. The resulting algorithm is well suited to sample from the joint smoothing distribution, but appears unnecessarily complex, computationally, for approximating additive smoothing functionals of the form (1).

In this paper, we study an SMC technique to smooth additive functionals based on a fixed-lag smoother presented by Kitagawa and Sato [11]. The method exploits the *forgetting properties* on the conditional hidden chain and is not affected by the degeneracy of the particle trajectories. Compared to Doucet *et al.* [7], computational requirements are marginal. Furthermore, we perform, under suitable regularity assumptions on the latent chain, a theoretical analysis of the behavior of the estimates obtained. It turns out that the L^p error and bias are upper bounded by quantities proportional to $n\log n/\sqrt{N}$ and $n\log n/N$, respectively, where N denotes the number of particles and n the number of observations.

In comparison, applying the results of Del Moral and Doucet ([4], Theorem 4) to a functional of type (1) provides a bound proportional to $n^2/\sqrt{N}$ on the L^p error for the standard trajectory-based particle smoother. Finally, we apply, for a noisily observed autoregressive model and the stochastic volatility model proposed by Hull and White [9], the technique to the *Monte Carlo EM* (MCEM) *algorithm* (Wei and Tanner [15]).

2. Particle approximation of additive functionals

2.1. The smoothing recursion

The *joint smoothing distribution* $\phi_{\nu,0:n|n}$ is the probability measure defined, for $A \in \mathcal{X}^{\otimes(n+1)}$, by

$$\phi_{\nu,0:n|n}[Y_{0:n}](A;\theta) \triangleq \mathbb{P}_\theta(X_{0:n} \in A|\mathcal{G}_n).$$

Under the assumptions above, the joint smoothing distribution has a density (for which we will use the same symbol) with respect to $\mu^{\otimes(n+1)}$ satisfying, for all $y_{0:k+1} \in \mathsf{Y}^{k+2}$, the recursion

$$\begin{aligned}&\phi_{\nu,0:k+1|k+1}[y_{0:k+1}](x_{0:k+1};\theta)\\ &\quad= \frac{\mathrm{L}_{\nu,k}(\theta;y_{0:k})}{\mathrm{L}_{\nu,k+1}(\theta;y_{0:k+1})} q_\theta(x_k,x_{k+1}) g_\theta(x_{k+1},y_{k+1}) \phi_{\nu,0:k|k}[y_{0:k}](x_{0:k};\theta). \qquad (2)\end{aligned}$$

For notational conciseness, we will omit the explicit dependence on the observations from the notation for the smoothing measure and replace $\phi_{\nu,0:k|k}[y_{0:k}](\cdot;\theta)$ by $\phi_{\nu,0:k|k}(\cdot;\theta)$.

Particle filtering, in its most basic form, consists of approximating the exact smoothing relations by propagating particle trajectories in the state space of the hidden chain. Given

a fixed sequence of observations, this is done according to the following scheme. In order to keep the notation simple, we fix the model parameters and omit θ from the notation throughout this part.

At time zero, N random variables $\{\xi_0^{N,i}; 1 \le i \le N\}$ are drawn from a common probability measure ς such that $\nu \ll \varsigma$. These *initial particles* are assigned the *importance weights* $\omega_0^{N,i} \triangleq W_0(\xi_0^{N,i})$, $1 \le i \le N$, where, for $x \in \mathsf{X}$, $W_0(x) \triangleq g(x, y_0)\,\mathrm{d}\nu/\mathrm{d}\varsigma(x)$, providing $\sum_{i=1}^N \omega_0^{N,i} f(\xi_0^{N,i}) / \sum_{i=1}^N \omega_0^{N,i}$ as an importance sampling estimate of $\phi_{\nu,0|0} f$ for $f \in \mathcal{B}_\mathrm{b}(\mathsf{X})$. Henceforth, the particle paths $\xi_{0:m}^{N,i} \triangleq [\xi_{0:m}^{N,i}(0), \ldots, \xi_{0:m}^{N,i}(m)]$, $1 \le i \le N$, are recursively updated according to the following procedure.

At time k, let $\{(\xi_{0:k}^{N,i}, \omega_k^{N,i}); 1 \le i \le N\}$ be a set of weighted particles approximating $\phi_{\nu,0:k|k}$, in the sense that $\sum_{i=1}^N \omega_k^{N,i} f(\xi_{0:k}^{N,i}) / \Omega_k^N$, with $\Omega_k^N \triangleq \sum_{i=1}^N \omega_k^{N,i}$ and $f \in \mathcal{B}_\mathrm{b}(\mathsf{X}^{k+1})$, is an estimate of the expectation $\phi_{\nu,0:k|k} f$. Then, an updated weighted sample $\{(\xi_{0:k+1}^{N,i}, \omega_{k+1}^{N,i}); 1 \le i \le N\}$, approximating the distribution $\phi_{\nu,0:k+1|k+1}$, is obtained by, first, simulating $\xi_{0:k+1}^{N,i} \sim R_k^\mathrm{p}(\xi_{0:k}^{N,i}, \cdot)$, where the kernel R_k^p is of type $R_k^\mathrm{p}(x_{0:k}, f) = \int_\mathsf{X} f(x_{0:k}, x_{k+1}) R_k(x_k, \mathrm{d}x_{k+1})$, with $f \in \mathcal{B}_\mathrm{b}(\mathsf{X}^{k+2})$ and each R_k being a Markov transition kernel. The new particles are simulated independently of each other and the special form of R_k^p implies that past particle trajectories are kept unchanged throughout this *mutation step*. A popular choice is to set $R_k \equiv Q$, yielding the so-called *bootstrap filter*; more sophisticated techniques involve proposals depending on the observed values (see Example 4.2). Second, when the observation $Y_{k+1} = y_{k+1}$ is available, the importance weights are updated according to $\omega_{k+1}^{N,i} = \omega_k^{N,i} W_{k+1}[\xi_{0:k+1}^{N,i}(k:k+1)]$, where, for $(x, x') \in \mathsf{X}^2$, $W_k(x, x') \triangleq g(x', y_k)\,\mathrm{d}Q(x, \cdot)/\mathrm{d}R_{k-1}(x, \cdot)(x')$. Now, for $f \in \mathcal{B}_\mathrm{b}(\mathsf{X}^{k+2})$, the self-normalized estimate $\phi_{\nu,0:k+1|k+1}^N f \triangleq \sum_{j=1}^N \omega_{k+1}^{N,j} f(\xi_{0:k+1}^{N,j}) / \Omega_{k+1}^N$ provides an approximation of $\phi_{\nu,0:k+1|k+1}$.

To prevent degeneracy, a *resampling mechanism* is introduced. In its simpler form, resampling amounts to drawing, conditionally independently, indices $I_k^{N,1}, \ldots, I_k^{N,N}$ from the set $\{1, \ldots, N\}$, multinomially with respect to the normalized weights $\omega_k^{N,j}/\Omega_k^N$, $1 \le j \le N$. Now, a new equally weighted sample $\{\hat{\xi}_{0:k}^{N,i}; 1 \le i \le N\}$ is constructed by setting $\hat{\xi}_{0:k}^{N,j} = \xi_{0:k}^{N,I_k^{N,j}}$. After the resampling procedure, the weights are all reset as $\omega_k^{N,i} = 1/N$, yielding another estimate, $\widehat{\phi}_{\nu,0:k|k}^N f \triangleq \sum_{i=1}^N f(\hat{\xi}_{0:k}^{N,i})/N$, of $\phi_{\nu,0:k|k}$. Note that the resampling mechanism might modify the whole trajectory of a certain particle, implying that, in general, for $m \le n$, $\xi_{0:n}^{N,i}(m) \ne \xi_{0:n+1}^{N,i}(m)$. The multinomial resampling method is not the only conceivable way to carry out the selection step (see e.g. Doucet *et al.* [6]).

Using the weighted samples $\{(\xi_{0:k}^{N,j}, \omega_k^{N,j}); 1 \le j \le N\}$, $0 \le k \le n$, produced under the parameter $\theta \in \Theta$, an approximation of $\gamma_{\theta,n} \triangleq \mathbb{E}_\theta[t_n(X_{0:n}) | \mathcal{G}_n]$ is obtained by constructing the estimators

$$\gamma_{\theta,n}^N = \frac{1}{\Omega_n^N} \sum_{j=1}^N \omega_n^{N,j} t_n(\xi_{0:n}^{N,j}) \quad \text{or} \quad \widehat{\gamma}_{\theta,n}^N = \frac{1}{N} \sum_{j=1}^N t_n(\hat{\xi}_{0:n}^{N,j}). \tag{3}$$

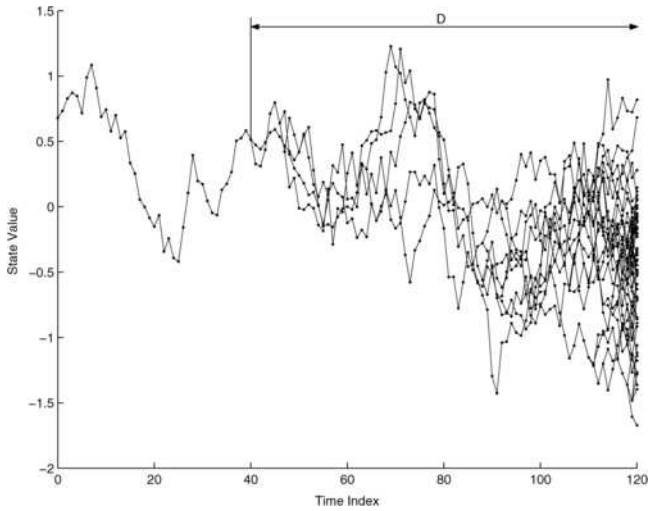

Figure 1. Typical particle trajectories for $N = 50$; see Section 4 for details regarding model and algorithm.

When the functional $\{t_n\}$ has the form given in (1), it is straightforward to verify that recording all of the particle trajectories is indeed not required to evaluate (3): upon defining $t_k^{N,i} \triangleq t_k(\xi_{0:k}^{N,j})$, we have, for $k \geq 1$,

$$t_{k+1}^{N,i} = \begin{cases} t_k^{N,i} + s_k[\xi_{0:k+1}^{N,i}(k:k+1)], & \text{if no resampling occurs,} \\ t_k^{N,I_{k+1}^i} + s_k[\hat{\xi}_{0:k+1}^{N,i}(k:k+1)], & \text{if resampling occurs.} \end{cases} \tag{4}$$

The recursion is initialized by $t_1^{N,i} = t_1(\xi_{0:1}^{N,i})$. In accordance with (3), γ_n^N is obtained as $\sum_{i=1}^N \omega_n^{N,i} t_n^{N,i} / \Omega_n^N$. Hence, for each particle $\xi_{0:k}^{N,i}$, we need only record its current position $\xi_{0:k}^{N,i}(k)$, weight $\omega_k^{N,i}$ and associated functional value $t_k^{N,i}$. Thus, the method necessitates only minor adaptations once the particle filter has been implemented.

As illustrated in Figure 1, as n increases, the path trajectories system collapses, and the estimators (3) are not reliable for sensible N values (see Doucet *et al.* [6], Kitagawa and Sato [11] and Andrieu and Doucet [1] for a discussion).

To cope with this drawback, we suggest the following method, based on a technique proposed by Kitagawa and Sato [11]. By the forgetting property of the time-reversed conditional hidden chain (Theorem 3.1), we expect that, for a large enough integer $\Delta_n \leq n - k$,

$$\mathbb{E}_\theta[s_k(X_{k:k+1})|\mathcal{G}_n] \approx \mathbb{E}_\theta[s_k(X_{k:k+1})|\mathcal{G}_{k+\Delta_n}], \tag{5}$$

yielding, with $k(\Delta_n) \triangleq (k + \Delta_n) \wedge n$,

$$\gamma_{\theta,n} = \mathbb{E}_\theta \left[\sum_{k=0}^{n-1} s_k(X_{k:k+1}) \Big| \mathcal{G}_n \right] \approx \sum_{k=0}^{n-1} \mathbb{E}_\theta [s_k(X_{k:k+1}) | \mathcal{G}_{k(\Delta_n)}].$$

The above relation suggests that waiting for all of the trajectories to collapse – as (4) implies – is not convenient. Instead, when the particle population N is sufficiently large so that (5) is valid for a lag Δ_n which may be far smaller than the typical collapsing time, one should apply the two approximations

$$\gamma_{\theta,n}^{N,\Delta_n} \triangleq \sum_{k=0}^{n-1} \sum_{j=1}^{N} \frac{\omega_{k(\Delta_n)}^{N,j}}{\Omega_{k(\Delta_n)}^{N}} s_k[\xi_{0:k(\Delta_n)}^{N,j}(k:k+1)], \tag{6}$$

$$\widehat{\gamma}_{\theta,n}^{N,\Delta_n} \triangleq \frac{1}{N} \sum_{k=0}^{n-1} \sum_{j=1}^{N} s_k[\xi_{0:k(\Delta_n)}^{N,j}(k:k+1)] \tag{7}$$

of $\gamma_{\theta,n}$. Although somewhat more involved than the standard approximation (3), the above lag-based approximation may be updated recursively by recording the recent history of the particles as well as the accumulated contribution of terms that will no longer get updated. Thus, apart from increased storage requirements, computing the lag-based approximation $\widehat{\gamma}_{\theta,n}^{N,\Delta_n}$ is clearly not, from a computational point of view, more demanding than computing $\widehat{\gamma}_{\theta,n}^{N}$.

3. Theoretical evaluation of the fixed-lag technique

To accomplish the robustification above, we need to specify the lag Δ_n and how this lag should depend on n. This is done by examining the quality of the estimates produced by the algorithm in terms of bias and L^p error. Of particular interest is how these errors are affected by the lag and whether it makes their dependence on n and N more favorable in comparison with the standard trajectory-based approach.

The validity of 5 is based on the assumption that the conditional hidden chains – in the forward as well as the backward directions – have *forgetting properties*, that is, the distributions of two versions of each chain starting at different initial distributions approach each other as time increases. This property depends on the following uniform ergodicity conditions on the model, which imply that forgetting occurs at a *geometrical* rate:

(A1) (i) $\sigma_- \triangleq \inf_{\theta\in\Theta} \inf_{x,x'\in\mathsf{X}} q_\theta(x,x') > 0$, $\sigma_+ \triangleq \sup_{\theta\in\Theta} \sup_{x,x'\in\mathsf{X}} q_\theta(x,x') < \infty$;
(ii) for all $y \in \mathsf{Y}$, $\sup_{\theta\in\Theta} \|g_\theta(\cdot,y)\|_{\mathsf{X},\infty} < \infty$, $\inf_{\theta\in\Theta} \int_{\mathsf{X}} g_\theta(x,y)\mu(\mathrm{d}x) > 0$.

Under (A1), we define

$$\rho \triangleq 1 - \frac{\sigma_-}{\sigma_+}. \tag{1}$$

We now define the Markov transition kernels that generate the conditional hidden chains. For any two transition kernels K and T from $(\mathsf{E}_1, \mathcal{E}_1)$ to $(\mathsf{E}_2, \mathcal{E}_2)$ and $(\mathsf{E}_2, \mathcal{E}_2)$ to $(\mathsf{E}_3, \mathcal{E}_3)$, respectively, we define the product transition kernel by $KT(x, A) \triangleq \int_{\mathsf{E}_2} T(z, A)K(x, \mathrm{d}z)$ for $x \in \mathsf{E}_1$ and $A \in \mathcal{E}_3$.

Introduce, for $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+2})$, $x_{0:k} \in \mathsf{X}^{k+1}$ and $y_{k+1} \in \mathsf{Y}$, the unnormalized pathwise transition kernel $L_k(x_{0:k}, f; \theta) \triangleq \int_{\mathsf{X}} f(x_{0:k+1}) g_\theta(x_{k+1}, y_{k+1}) Q_\theta(x_k, \mathrm{d}x_{k+1})$. Assumption (A1) makes this integral well defined for all $k \geq 0$. We will often consider compositions

$$L_k \cdots L_m(x_{0:k}, f; \theta) = \int_{\mathsf{X}^{m-k+1}} f(x_{0:m+1}) \prod_{i=k}^{m} [g_\theta(x_{i+1}, y_{i+1}) Q_\theta(x_i, \mathrm{d}x_{i+1})]$$

with $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{m+2})$, $x_{0:k} \in \mathsf{X}^{k+1}$ and $y_{0:k} \in \mathsf{Y}^{m-k+1}$, and it is clear that, for all $k \leq m$, the function $L_k \cdots L_m(x_{0:k}, \mathsf{X}^{m+2}; \theta)$ depends only on x_k. Thus, a version of this function comprising only the last component is well defined and we write $L_k \cdots L_m(x_k, \mathsf{X}^{m+2}; \theta)$ in this case. For $k > m$, we set $L_k \cdots L_m \equiv \mathrm{Id}$. Using this notation and given $n \geq 0$, the *forward smoothing kernels* given by, for $k \geq 0$, $x_k \in \mathsf{X}$ and $A \in \mathcal{X}$, $\mathrm{F}_{k|n}(x_k, A; \theta) \triangleq \mathbb{P}_\theta(X_{k+1} \in A | X_k = x_k, \mathcal{G}_n)$, can, for indices $0 \leq k < n$ and $y_{k+1} \in \mathsf{Y}$, be written as

$$\begin{aligned} &\mathrm{F}_{k|n}(x_k, A; \theta) \\ &\quad = \begin{cases} \displaystyle\int_A \frac{g_\theta(x_{k+1}, y_{k+1}) L_{k+1} \cdots L_{n-1}(x_{k+1}, \mathsf{X}^{n+1}; \theta) Q_\theta(x_k, \mathrm{d}x_{k+1})}{L_k \cdots L_{n-1}(x_k, \mathsf{X}^{n+1}; \theta)}, \\ \qquad\qquad \text{for } \ 0 \leq k < n, \\ Q_\theta(x_k, A), \qquad \text{for } \ k \geq n. \end{cases} \end{aligned} \tag{2}$$

Analogously, for the time-reversed conditional hidden chain, we consider the *backward smoothing kernels* defined by $\mathrm{B}_{\nu,k|n}(x_{k+1}, A; \theta) \triangleq \mathbb{P}_\theta(X_k \in A | X_{k+1} = x_{k+1}, \mathcal{G}_n)$, where $k \geq 0$, $x_{k+1} \in \mathsf{X}$ and $A \in \mathcal{X}$. Note that $\mathrm{B}_{\nu,k|n}$ depends on the initial distribution of the latent chain. The backward kernel can be expressed as

$$\begin{aligned} &\mathrm{B}_{\nu,k|n}(x_{k+1}, A; \theta) \\ &\quad = \begin{cases} \dfrac{\int_A q_\theta(x_k, x_{k+1}) \phi_{\nu,k}(\mathrm{d}x_k; \theta)}{\int_{\mathsf{X}} q_\theta(x'_k, x_{k+1}) \phi_{\nu,k}(\mathrm{d}x'_k; \theta)}, & \text{for } 0 \leq k \leq n, \\ \dfrac{\int_A \int_{\mathsf{X}} q_\theta(x_k, x_{k+1}) q_\theta^{k-n}(x_n, x_k) \phi_{\nu,n}(\mathrm{d}x_n; \theta) \mu(\mathrm{d}x_k)}{\int_{\mathsf{X}} q_\theta^{k-n+1}(x'_n, x_{k+1}) \phi_{\nu,n}(\mathrm{d}x'_n; \theta)}, & \text{for } k > n, \end{cases} \end{aligned}$$

where, for $m \geq 1$, q_θ^m denotes the density of the m-step kernel Q_θ^m.

The following theorem (see Del Moral [3], page 143), stating geometrical ergodicity of the forward and backward chains, is instrumental for the developments which are to follow.

Theorem 3.1. *Assume* (A1) *and let* ρ *be defined in (1). Then, for all* $k \geq m \geq 0$*, all* $\theta \in \Theta$*, all probability measures* ν_1*,* ν_2 *on* $\mathcal{X}$ *and all* $y_{0:n} \in \mathsf{Y}^{n+1}$*,*

$$\|\nu_1 \mathrm{F}_{m|n} \cdots \mathrm{F}_{k|n}(\cdot;\theta) - \nu_2 \mathrm{F}_{m|n} \cdots \mathrm{F}_{k|n}(\cdot;\theta)\|_{\mathrm{TV}} \leq \rho^{k-m+1},$$

$$\|\nu_1 \mathrm{B}_{\nu,k|n} \cdots \mathrm{B}_{\nu,m|n}(\cdot;\theta) - \nu_2 \mathrm{B}_{\nu,k|n} \cdots \mathrm{B}_{\nu,m|n}(\cdot;\theta)\|_{\mathrm{TV}} \leq \rho^{k-m+1}.$$

Assumption (A1) typically requires that X is a compact set, but some very recent papers (Douc et al. [5], Kleptsyna and Veretennikov [12]) provide results that establish geometric forgetting under considerably weaker assumptions. Applying these results within our framework would, however, make the analysis far more complicated since the provided bounds are uniform neither in the observations nor the initial distributions.

3.1. Main results

For the sake of simplicity, let us assume that multinomial resampling is applied at every iteration. Moreover, let the observations used by the particle filter be generated by a state space model with kernel, measurement density and initial distribution $\bar{Q}$, $\bar{g}$ and $\bar{\nu}$, respectively. We stress that $\bar{Q}$ and $\bar{g}$ are not assumed to belong to the parametric family $\{(Q_\theta, g_\theta); \theta \in \Theta\}$. Using these observed values as input, the evolution of the particle cloud follows the usual dynamics $(Q_\theta, g_\theta, \nu, \theta \in \Theta)$ and, in this setting, it is easily verified that the process $\{Z_k; k \geq 0\}$, with $Z_k \triangleq [\xi_{0:k}^{N,1}(k-1:k), \ldots, \xi_{0:k}^{N,N}(k-1:k), X_k, Y_k]$, is a Markov chain on $\mathsf{X}^{2N+1} \times \mathsf{Y}$. We denote by $\bar{\mathbb{P}}_\theta^N$ and $\bar{\mathbb{E}}_\theta^N$ the law of this chain and the associated expectation, respectively, and define the filtration $\{\mathcal{F}_k^N; k \geq 0\}$ by $\mathcal{F}_{k+1}^N \triangleq \mathcal{F}_k^N \vee \sigma(\xi_{0:k+1}^{N,1}, \ldots, \xi_{0:k+1}^{N,N})$ with $\mathcal{F}_0^N \triangleq \sigma(\xi_0^{N,1}, \ldots, \xi_0^{N,N})$. The marginal of $\bar{\mathbb{P}}_\theta^N$ with respect to $\{(X_k, Y_k); k \geq 0\}$ and the associated expectation are denoted by $\bar{\mathbb{P}}$ and $\bar{\mathbb{E}}$, respectively. For any integer $p \geq 1$, random variable $V \in \mathsf{L}^p(\bar{\mathbb{P}}_\theta^N)$ and sub-σ-algebra $\mathcal{A} \subseteq \sigma(\{Z_k; k \geq 0\})$ we define the conditional L^p norm $\|V\|_{p|\mathcal{A}} \triangleq (\bar{\mathbb{E}}_\theta^N[|V|^p|\mathcal{A}])^{1/p}$.

(A2) For all $k \geq 0$, $\theta \in \Theta$ and $y_k \in \mathsf{Y}$, $\|W_k(\cdot;\theta)\|_{\mathsf{X}^2,\infty} < \infty$.

Remark 3.2. In case of the bootstrap particle filter, for which $R_{\theta,k} \equiv Q_\theta$, assumption (A2) is implied by assumption (A1). The same is true for the so-called *optimal kernel* used in Example 4.2.

Theorem 3.3. *Assume* (A1) *and* (A2)*. There then exist universal constants* B_p *and* B*,* B_p *depending only on* p*, such that the following holds true for all* $n \geq 0$*,* $\theta \in \Theta$*,* $\Delta_n \geq 0$ *and* $N \geq 1$*:*

(i) *for all* $p \geq 2$*,*

$$\|\widehat{\gamma}_{\theta,n}^{N,\Delta_n} - \gamma_{\theta,n}\|_{p|\mathcal{G}_n} \leq 2\rho^{\Delta_n} \sum_{k=0}^{n-\Delta_n} \|s_k\|_{\mathsf{X}^2,\infty}$$

$$+\frac{B_p}{\sqrt{N}(1-\rho)}\sum_{k=0}^{n-1}\|s_k\|_{\mathsf{X}^2,\infty}\left[\frac{1}{\sigma_-}\sum_{m=1}^{k(\Delta_n)}\frac{\|W_m(\cdot;\theta)\|_{\mathsf{X}^2,\infty}\rho^{0\vee(k-m)}}{\mu g_\theta(Y_m)}\right.$$
$$\left.+\frac{\|W_0(\cdot;\theta)\|_{\mathsf{X},\infty}\rho^k}{\nu g_\theta(Y_0)}+1\right];$$

(ii)

$$|\bar{\mathbb{E}}_\theta^N[\widehat{\gamma}_{\theta,n}^{N,\Delta_n}|\mathcal{G}_n]-\gamma_{\theta,n}|$$
$$\le 2\rho^{\Delta_n}\sum_{k=0}^{n-\Delta_n}\|s_k\|_{\mathsf{X}^2,\infty}$$
$$+\frac{B}{N(1-\rho)^2}\sum_{k=0}^{n-1}\|s_k\|_{\mathsf{X}^2,\infty}\left[\frac{1}{\sigma_-^2}\sum_{m=1}^{k(\Delta_n)}\frac{\|W_m(\cdot;\theta)\|_{\mathsf{X}^2,\infty}^2\rho^{0\vee(k-m)}}{\{\mu g_\theta(Y_m)\}^2}\right.$$
$$\left.+\frac{\|W_0(\cdot;\theta)\|_{\mathsf{X},\infty}^2\rho^k}{\{\nu g_\theta(Y_0)\}^2}\right].$$

For the purpose of illustrating these bounds, assume that we are given a set $\{y_k;k\ge 0\}$ of fixed observations and that all $\|s_k\|_{\mathsf{X}^2,\infty}$, as well as all fractions $\|W_k(\cdot;\theta)\|_{\mathsf{X}^2,\infty}/\mu g_\theta(y_k)$, are uniformly bounded in k. We then conclude that increasing the lag with n as $\Delta_n=\lceil c\log n\rceil$, $c>-1/\log\rho$, will imply that $n\rho^{\Delta_n}$ tends to zero as n goes to infinity, leading to an error which is dominated by the variability due to the particle filter (the second term of the bound in Theorem 3.3(i)) and upper bounded by a quantity proportional to

$$\frac{1}{\sqrt{N}}\sum_{k=0}^{n-1}\left[\sum_{m=1}^{k+\lceil c\log n\rceil}\rho^{0\vee(k-m)}+1\right]\le\frac{n}{\sqrt{N}}\left(\frac{1}{1-\rho}+1+\lceil c\log n\rceil\right),$$

that is, of order $n\log n/\sqrt{N}$. Note the dependence on the mixing coefficient ρ of this rate. In contrast, setting $\Delta_n=n$, that is, using the direct full-path approximation, would result in a stochastic error which is upper bounded by a quantity proportional to $n^2/\sqrt{N}$.

3.2. Extension to randomly varying observations

As mentioned, all results presented above concern smoothing distribution approximations produced by the particle filter algorithm *conditionally* on a given sequence of observations. In this section, we extend these results to the case of a randomly varying observation sequence.

For the bounds presented in Theorem 3.3, the conditioning on $\mathcal{G}_n$ can be removed by introducing additional model assumptions. In the following, we suppose that $\nu\ll\mu$ and that the resulting Radon–Nikodým derivative satisfies $(\mathrm{d}\nu/\mathrm{d}\mu)_-\triangleq\inf_{x\in\mathsf{X}}\mathrm{d}\nu/\mathrm{d}\mu(x)>0$.

(A3) Let t_n be given by (1). For $p \geq 2$, $\ell \geq 1$ and $\theta \in \Theta$, there exists a constant $a_{p,\ell}(t_n;\theta) \in \mathbb{R}^+$ such that

$$\max\left\{\bar{\mathbb{E}}\left[\frac{\|W_k(\cdot;\theta)\|_{\mathsf{X},\infty}^p \|s_i\|_{\mathsf{X}^2,\infty}^\ell}{\{\mu g_\theta(Y_k)\}^p}\right], \bar{\mathbb{E}}[\|s_i\|_{\mathsf{X}^{n+1},\infty}^\ell]; 0 \leq k \leq n, 0 \leq i \leq n-1\right\} \leq a_{p,\ell}(t_n;\theta).$$

Proposition 3.4. *Assume* (A1) *and* (A2). *There then exist universal constants* B_p *and* B, B_p *depending only on* p, *such that the following holds true for all* $N \geq 1$*:*

(i) *if assumption* (A3) *is satisfied for* $\ell = p \geq 2$ *and* $\theta \in \Theta$, *then*

$$\|\widehat{\gamma}_{\theta,n}^{N,\Delta_n} - \gamma_{\theta,n}\|_p \leq 2a_{p,p}^{1/p}(t_n;\theta)\rho^{\Delta_n}(n-\Delta_n+1) + \frac{B_p a_{p,p}^{1/p}(t_n;\theta)}{\sqrt{N}(1-\rho)}\left\{\frac{\Delta_n(n+1)}{\sigma_-} + n\left[\frac{1}{\sigma_-(1-\rho)} + \frac{1}{(\mathrm{d}\nu/\mathrm{d}\mu)_-^2} + 1\right]\right\};$$

(ii) *if assumption* (A3) *is satisfied for* $p = 2$, $\ell = 1$ *and* $\theta \in \Theta$, *then*

$$|\bar{\mathbb{E}}_\theta^N[\widehat{\gamma}_{\theta,n}^{N,\Delta_n} - \gamma_{\theta,n}]| \leq 2a_{2,1}(t_n;\theta)\rho^{\Delta_n}(n-\Delta_n+1) + \frac{Ba_{2,1}(t_n;\theta)}{N(1-\rho)^2}\left\{\frac{\Delta_n(n+1)}{\sigma_-^2} + n\left[\frac{1}{\sigma_-^2(1-\rho)} + \frac{1}{(\mathrm{d}\nu/\mathrm{d}\mu)_-^2}\right]\right\}.$$

The proof of this result is given in Section A.2.

Remark 3.5. In the case of a compact state space X, assumption (A3) implies only limited additional restrictions on the state space model. In fact, for a large class of models, assumption (A3) follows as a direct consequence of assumption (A1).

4. Applications to maximum likelihood estimation

We now return to the computation of the maximum likelihood estimator. In the following, we consider models for which the set of complete data likelihood functions is an *exponential family*, that is, for all $\theta \in \Theta$ and $n \geq 0$, the joint density of $(X_{0:n}, Y_{0:n})$ is of the form $\exp[\langle \psi(\theta), S_n(x_{0:n})\rangle - c(\theta)]h(x_{0:n})$. Here, ψ and the sufficient statistics S_n are $\mathbb{R}^{d_s}$-valued functions on Θ and X^{n+1}, respectively, c is a real-valued function on θ and h is a real-valued non-negative function on X^{n+1}. By $\langle \cdot, \cdot \rangle$ we denote the scalar product. All of these functions may depend on the observed values $y_{0:n}$, even though this is expunged from the notation.

If the complete data likelihood function is of the particular form above and the expectation $\phi_{\nu,0:n|n}(S_n;\theta)$ is finite for all $\theta \in \Theta$, then the intermediate quantity of EM can be written as (up to quantities which do not depend on θ) $\mathcal{Q}(\theta;\theta') = \langle \psi(\theta), \phi_{\nu,0:n|n}(S_n;\theta')\rangle - c(\theta)$.

Note, finally that, as mentioned in the Introduction, a typical element $S_{n,m}(x_{0:n})$, $1 \leq m \leq d_s$, of the vector $S_n(x_{0:n})$ is an additive functional $S_{n,m}(x_{0:n}) = \sum_{k=0}^{n-1} s_{n,m}^{(k)}(x_{k:k+1})$ so that $\phi_{\nu,0:n|n}(S_n;\theta')$ can be estimated using either (6) or (7). Denoting by $\widehat{S}_n$ such an estimate, we may approximate the intermediate quantity by

$$\widehat{\mathcal{Q}}^N(\theta;\theta') = \langle \psi(\theta), \widehat{S}_n \rangle - c(\theta).$$

In the next step – referred to as the M-step – $\widehat{\mathcal{Q}}^N(\theta;\theta')$ is maximized with respect to θ, providing a new parameter estimate. This procedure is repeated recursively given an initial guess $\widehat{\theta}_0$.

As an illustration, we consider the problem of inference in a noisily observed AR(1) model and the stochastic volatility (SV) model. None of these examples satisfy assumption (A1); however, geometric ergodicity for the models in question can be established using bounds presented by Douc et al. [5]. Although these bounds are somewhat more involved than those presented in Theorem 3.1 (e.g., the former depend on the initial distributions and the observations), we may, nevertheless, expect that the conclusion reached in Section 3, that is, that the error of the fixed-lag approximation is controlled by a lag of order $\log n$, still applies. The situation is complicated, however, by the fact that the mixing rates depend on the observations and are uniform only under the expectation operator. In other words, there may be occasional outcomes for which mixing is poor, even if the average performance of the system is satisfactory.

Example 4.1 (SMCEM for noisily observed AR(1) model). We consider the state space model

$$X_{k+1} = aX_k + \sigma_w W_{k+1},$$
$$Y_k = X_k + \sigma_v V_k$$

with $\{W_k; k \geq 1\}$ and $\{V_k; k \geq 0\}$ being mutually independent sets of standard normal distributed variables such that W_{k+1} is independent of (X_i, Y_i), $0 \leq i \leq k$, and V_k is independent of X_k, (X_i, Y_i), $0 \leq i \leq k-1$. The initial distribution is chosen to be a diffuse prior so that $\phi_{\nu,0|0}$ is $\mathcal{N}(y_0, \sigma_v^2)$. Throughout the experiment, we use a fixed sequence of observations produced by simulation under the parameters $a^* = 0.98$, $\sigma_w^* = 0.2$ and $\sigma_v^* = 1$. In this case, $\psi(\theta) = [1/2\sigma_w^2, -a/\sigma_w^2, a^2/(2\sigma_w^2), 1/(2\sigma_v^2)]$ and the components of the $\mathbb{R}^4$-valued function $x_{0:n} \mapsto S_n(x_{0:n})$ are given by $S_{n,1}(x_{0:n}) \triangleq \sum_{k=1}^{n-1} x_k^2$, $S_{n,2}(x_{0:n}) \triangleq \sum_{k=0}^{n-1} x_k x_{k+1}$, $S_{n,3}(x_{0:n}) \triangleq \sum_{k=0}^{n} x_k^2$ and $S_{n,4}(x_{0:n}) \triangleq \sum_{k=0}^{n}(y_k - x_k)^2$. Furthermore, up to terms not depending on parameters, $c(\theta) = n\log(\sigma_w^2)/2 + (n+1)\log(\sigma_v^2)/2$. In this setting, one step of the MCEM algorithm is carried out in the following way. Having produced an estimate $\widehat{\theta}^{i-1}$ of the parameters $\theta = (a, \sigma_w^2, \sigma_v^2)$ at the previous iteration, we compute an approximation $\widehat{S}_n = (\widehat{S}_{n,1}, \widehat{S}_{n,2}, \widehat{S}_{n,3}, \widehat{S}_{n,4})$ of $\phi_{\nu,0:n|n}(S_n;\widehat{\theta}^{i-1})$ using the particle filter and update the parameters according to

$$\widehat{a}^i = \frac{\widehat{S}_{n,2}}{\widehat{S}_{n,1}}, \qquad (\widehat{\sigma}_w^i)^2 = \frac{1}{n}(\widehat{S}_{n,3} - \widehat{a}^i \widehat{S}_{n,2}), \qquad (\widehat{\sigma}_v^i)^2 = \frac{\widehat{S}_{n,4}}{n+1}.$$

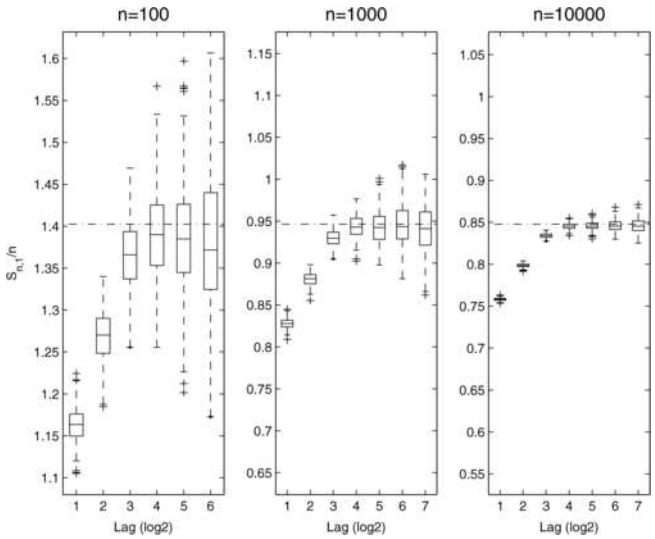

Figure 2. Boxplots of estimates of $\phi_{\nu,0:n|n}S_{n,1}/n$, produced with the fixed-lag technique, for the noisily observed AR(1) model in Example 4.1.

We simulated, for each $n = 100$, 1000, 10,000 observations, 1000 SMC estimates of $\phi_{\nu,0:n|n}S_1$ using the fixed-lag smoothing technique for the parameter values $a = 0.8$, $\sigma_w = 0.5$ and $\sigma_v = 2$. Here, the standard bootstrap particle filter with systematic resampling was used, with $R_k \equiv Q$ for all $k \geq 0$. The dotted lines indicate the exact expected values, obtained by means of disturbance smoothing. To study the bias-variance trade-off – discussed in detail in the previous section – of the method, we used six different lags for each n and a constant particle population size $N = 1000$. The result is displayed in Figure 2, from which it is evident that the bias is controlled for a size of lag that increases approximately logarithmically with n. In particular, from the plot, we deduce that an optimal outcome is gained when lags of size 2^4, 2^4 and 2^5 are used for n being 100, 1000 and 10,000, respectively.

When the lag is sufficiently large so that we can ignore the term of the bias which is deduced from forgetting arguments (being roughly of magnitude $n\rho^{\Delta_n}$), increasing the lag further exclusively leads to an increase of variance, as well as bias, of the estimates; compare the two last boxes of each plot. This is completely in accordance with the theoretical results of Section 2. Note that the scale on the y-axis is the same for the three panels, although the y-axis has been shifted in each panel due to the fact that the value of the normalized smoothed statistic evolves as the number of observations increases.

In Figure 3, we again report the cases $n = 100$, 1000, 10,000 observations and compare the basic approximation strategy (4) with the one based on fixed-lag smoothing with suitable lags. Guided by the plots of Figure 2 and the theory developed in the previous section, we choose the lags 2^4, 2^4 and 2^5, respectively. The number of particles was set to 1000 for all n. It is obvious that fixed-lag smoothing drastically reduces the variance

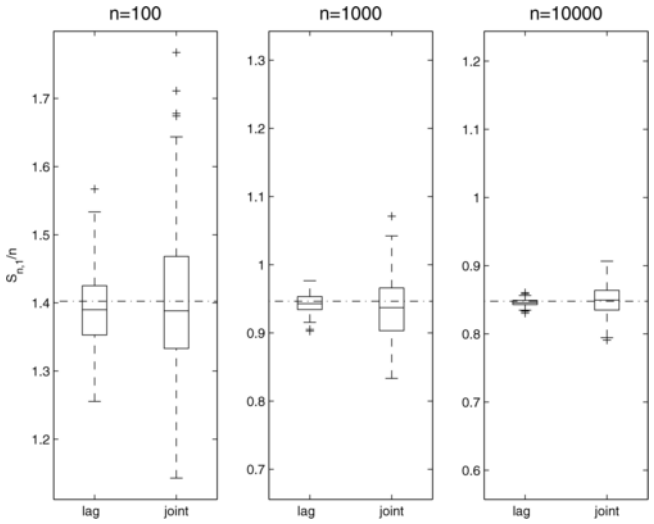

Figure 3. Boxplots of estimates of $\phi_{\nu,0:n|n} S_{n,1}/n$, produced by means of both the fixed-lag technique and standard trajectory-based smoothing, for the noisily observed AR(1) model in Example 4.1. Each box is based on 200 estimates, and the size of the particle population was $N = 1000$ for all cases.

without significantly raising the bias. As in the previous figure, dotted lines indicate exact values. As expected, the bias of the two techniques increases with n since the number of particles is held constant.

Example 4.2 (SMCEM for the stochastic volatility (SV) model). In the discrete-time case, the canonical version of the SV model (Hull and White [9], Jacquier *et al.* [10]) is given by the two relations

$$X_{k+1} = \alpha X_k + \sigma \epsilon_{k+1},$$
$$Y_k = \beta \exp(X_k/2)\varepsilon_k$$

with $\{\epsilon_k; k \geq 1\}$ and $\{\varepsilon_k; k \geq 0\}$ being mutually independent sets of standard normal distributed variables such that W_{k+1} is independent of (X_i, Y_i), $0 \leq i \leq k$, and V_k is independent of X_k, (X_i, Y_i), $0 \leq i \leq k-1$.

To use the SV model in practice, we need to estimate the parameters $\theta = (\beta, \alpha, \sigma)$. Throughout this example, we will use a sequence of data obtained by simulation under the parameters $\beta^* = 0.63$, $\alpha^* = 0.975$ and $\sigma^* = 0.16$. These parameters are consistent with empirical estimates for daily equity return series and are often used in simulation studies. In conformity with Example 4.1, we assume that the latent chain is initialized by an improper diffuse prior. The SV model is within the scope of exponential families, with $\psi(\theta) = [-\alpha^2/(2\sigma^2), -1/(2\sigma^2), \alpha/\sigma^2, -1/(2\beta^2)]$ and components of $S_n(x_{0:n})$ given by $S_{n,1}(x_{0:n}) \triangleq \sum_{k=0}^{n-1} x_k^2$, $S_{n,2}(x_{0:n}) \triangleq \sum_{k=1}^{n} x_k^2$, $S_{n,3}(x_{0:n}) \triangleq \sum_{k=1}^{n} x_k x_{k-1}$ and

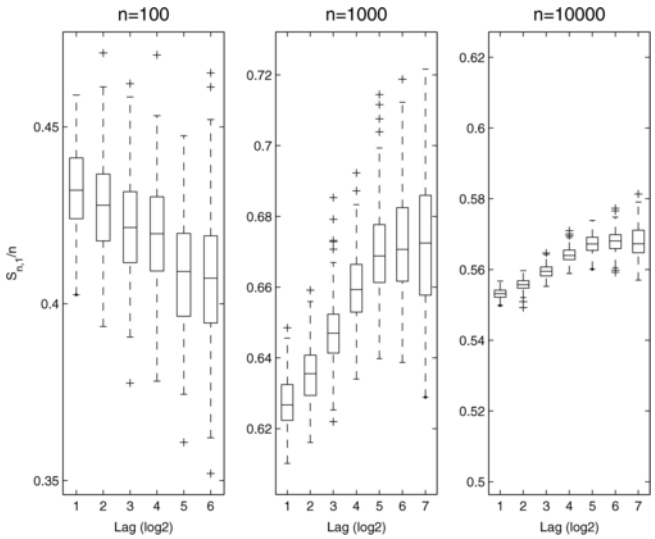

Figure 4. Boxplots of estimates of $\phi_{\nu,0:n|n}S_{n,1}/n$, produced with the fixed-lag technique, for the SV model in Example 4.2. Each box is based on 200 estimates and the size of the particle population was set to $N = 1000$ in all cases.

$S_{n,4}(x_{0:n}) \triangleq \sum_{k=0}^{n} y_k \exp(-x_k)$. In addition, up to terms not depending on parameters, $c(\theta) = (n+1)\log(\beta^2)/2 + (n+1)\log(\sigma^2)/2$.

Let $\widehat{S}_n = (\widehat{S}_{n,1}, \widehat{S}_{n,2}, \widehat{S}_{n,3}, \widehat{S}_{n,4})$ be a particle approximation of $\phi_{\nu,0:n|n}(S_n; \widehat{\theta}^{i-1})$. To apply the Monte Carlo EM algorithm to the SV model is not more involved than for the autoregressive model in Example 4.1. In fact, the updating formulas appear to be completely analogous:

$$\widehat{\alpha}^i = \frac{\widehat{S}_{n,3}}{\widehat{S}_{n,1}}, \qquad (\widehat{\sigma}^i)^2 = \frac{1}{n}(\widehat{S}_{n,2} - \widehat{\alpha}^i \widehat{S}_{n,3}), \qquad (\widehat{\beta}^i)^2 = \frac{\widehat{S}_{n,4}}{n+1}.$$

As proposal kernel R_k, we use an approximation, used by Cappé *et al.* ([2], Example 7.2.5) and inspired by Pitt and Shepard [13], of the so-called *optimal kernel*, that is, the conditional density of X_{k+1} given *both* X_k and Y_{k+1}.

We repeat the numerical investigations of Example 4.1. The resulting approximation of $\phi_{\nu,0:n|n}S_{n,1}$, displayed in Figure 4, behaves similarly. Here, again, we observe that moderate values of the lag Δ are sufficient to suppress the bias.

We finally compare the SMCEM parameter estimates obtained with the fixed-lag approximation and with the standard trajectory-based approximation on a simulated dataset of length $n = 5000$. Note that for the SMCEM procedure to converge to the MLE, it is necessary to increase the number of simulations that are performed as we progress through the EM iterations. We follow the recommendation of Fort and Moulines [8] and start by running 150 iterations of the Monte Carlo EM procedure with the number of

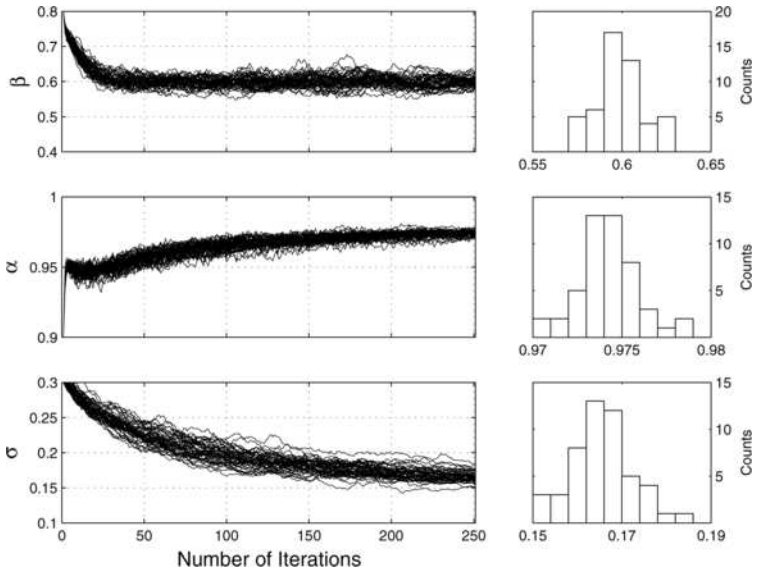

Figure 5. SMCEM parameter estimates of β, α and σ from $n = 5000$ observations using the standard trajectory-based smoothing approximation. Each plot overlays 50 realizations of the particle simulations; the histograms pertain to the final (250th) SMCEM iteration.

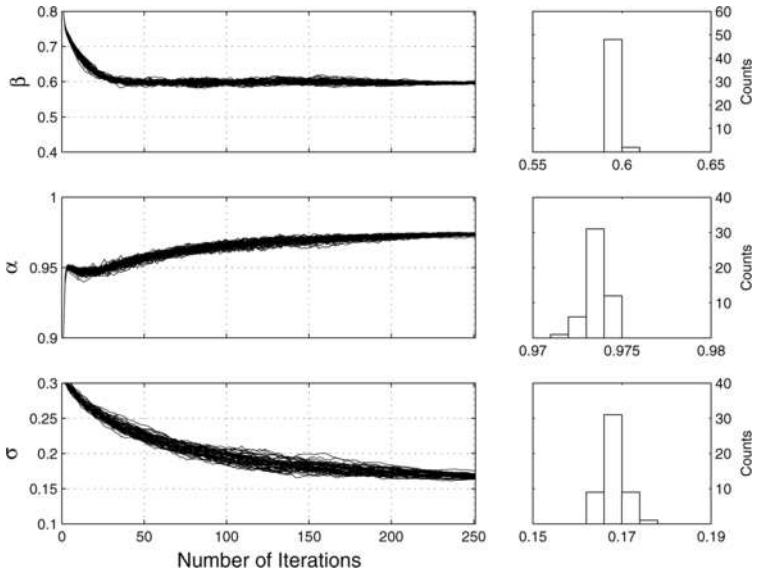

Figure 6. SMCEM parameter estimates of β, α and σ from $n = 5000$ observations using the fixed-lag smoothing approximation with $\Delta = 40$. Each plot overlays 50 realizations of the particle simulations; the histograms pertain to the final (250th) SMCEM iteration.

particles set at $N = 100$. For the subsequent 100 iterations, the number of particles increases at a quadratic rate with a final value (for the 250th Monte Carlo EM iteration) equal to $N = 1600$. The cumulative number of simulations performed during the 250 SMCEM iterations is equal to 75,000 (times the length of the observation sequence), which is quite moderate for a Monte Carlo-based optimization method. In Figures 5 and 6, we display the superimposed trajectories of parameter estimates for 50 realizations of the particles, together with histograms of the final estimates (at iteration 250) when using, respectively, the trajectory-based approximation (in Figure 5) and the fixed-lag approximation with $\Delta = 40$ (in Figure 6). Not surprisingly, the fact that the particle simulations are iterated for several successive values of the parameter estimates only amplifies the differences observed so far. With the fixed-lag approximation, the standard deviation of the final SMCEM parameter estimate is divided by a factor of seven for β, and three for α and σ, which is quite impressive in the context of Monte Carlo methods: to achieve the same accuracy with the trajectory-based approximation, one would need about ten times more particles to compensate for the higher simulation variance. Table 1 shows that the fixed-lag approximation (third row) indeed remains more reliable than the trajectory-based approximation, even when the latter is computed from ten times more particles (second row). Note that, for the trajectory-based approximation, multiplying the number of particles by ten does not reduce the standard deviation of the estimates as much as expected from the asymptotic theory. This is certainly due to the moderate number of particles used in the baseline setting, as we start from $N = 100$ particles during the first SMCEM iterations and terminate with $N = 1600$.

Appendix A: Proofs

A.1. Proof of Theorem 3.3

The proof of Theorem 3.3 partly comprises the geometric ergodicity of the time-reversed conditional hidden chain (Theorem 3.1), partly the next proposition. In the following, we omit θ from the notation for brevity. Moreover, let $\mathcal{C}_i(\mathsf{X}^{n+1})$ be the set of bounded

Table 1. Mean and standard deviation of SMCEM parameter estimates at the 250th iteration (estimated from 50 independent runs)

Smoothing algorithm	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\sigma}$
Trajectory-based,	0.5991	0.9742	0.1659
with 75,000 total simulations	std. 0.0136	std. 0.0019	std. 0.0070
Trajectory-based,	0.5990	0.9739	0.1666
with 750,000 total simulations	std. 0.0045	std. 0.0011	std. 0.0043
Fixed-lag,	0.5962	0.9735	0.1682
with 75,000 total simulations	std. 0.0019	std. 0.0006	std. 0.0024

measurable functions f on X^{n+1}, possibly depending on $Y_{0:n}$, of type $f(x_{0:n}) = \bar{f}(x_{i:n})$ for some function $\bar{f}: \mathsf{X}^{n-i+1} \to \mathbb{R}$.

Proposition A.1. *Assume* (A1) *and* (A2)*, and let* $f \in \mathcal{C}_i(\mathsf{X}^{n+1})$*,* $0 \le i \le n$*. There then exist universal constants* B_p *and* B*,* B_p *depending only on* p*, such that the following holds for all* $N \ge 1$*:*

(i) *for all* $p \ge 2$*,*

$$\|\widehat{\phi}^N_{\nu,0:n|n} f_i - \phi_{\nu,0:n|n} f_i\|_{p|\mathcal{G}_n} \le \frac{B_p \|f_i\|_{\mathsf{X}^{n+1},\infty}}{\sqrt{N}(1-\rho)} \left[\frac{1}{\sigma_-} \sum_{k=1}^{n} \frac{\|W_k\|_{\mathsf{X}^2,\infty} \rho^{0 \vee (i-k)}}{\mu g(Y_k)} + \frac{\|W_0\|_{\mathsf{X},\infty} \rho^i}{\nu g_0} + 1 \right];$$

(ii)

$$|\bar{\mathbb{E}}^N[\widehat{\phi}^N_{\nu,0:n|n} f_i | \mathcal{G}_n] - \phi_{\nu,0:n|n} f_i| \le \frac{B \|f_i\|_{\mathsf{X}^{n+1},\infty}}{N(1-\rho)^2} \left[\frac{1}{\sigma_-^2} \sum_{k=1}^{n} \frac{\|W_k\|^2_{\mathsf{X}^2,\infty} \rho^{0 \vee (i-k)}}{\{\mu g(Y_k)\}^2} + \frac{\|W_0\|^2_{\mathsf{X},\infty} \rho^i}{\{\nu g(Y_0)\}^2} \right].$$

To prove Proposition A.1, we need some preparatory lemmas and definitions. In accordance with the mutation-selection procedure presented in Section 2, we have, for $k \ge 1$, $A \in \mathcal{X}^{\otimes(k+1)}$ and $i \in \{1, \dots, N\}$, that

$$\begin{aligned} &\bar{\mathbb{P}}^N(\xi^{N,j}_{0:k} \in A | \mathcal{G}_k \vee \mathcal{F}^N_{k-1}) \\ &= \sum_{j=1}^{N} \bar{\mathbb{P}}^N(I^{N,i}_{k-1} = j | \mathcal{G}_k \vee \mathcal{F}^N_{k-1}) \bar{\mathbb{P}}^N(\xi^{N,j}_{0:k} \in A | I^{N,i}_{k-1} = j, \mathcal{G}_k \vee \mathcal{F}^N_{k-1}) \\ &= \sum_{j=1}^{N} \frac{\omega^{N,j}_{k-1}}{\Omega^N_{k-1}} R^{\mathrm{p}}_{k-1}(\xi^{N,j}_{0:k-1}, A). \end{aligned}$$

That is, conditional on $\mathcal{F}^N_{k-1}$, the swarm $\{\xi^{N,i}_{0:k}; 1 \le i \le N\}$ of mutated particles at time k is obtained by sampling N independent and identically distributed particles from the measure

$$\eta^N_k \triangleq \phi^N_{\nu,0:k-1|k-1} R^{\mathrm{p}}_{k-1}. \tag{A.1}$$

Using this, define, for $A \in \mathcal{X}^{\otimes(k+1)}$,

$$\mu^N_{k|n}(A) \triangleq \int_A \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k}(x_{0:k}) \eta^N_k(\mathrm{d}x_{0:k}), \tag{A.2}$$

where the Radon–Nikodým derivative is given by, for $x_{0:k} \in \mathsf{X}^{k+1}$,

$$\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(x_{0:k}) \triangleq \frac{W_k(x_{k-1:k})L_k \cdots L_{n-1}(x_{0:k}, \mathsf{X}^{n+1})}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})}.$$

In addition, for $A \in \mathcal{X}$, let

$$\mu_{0|n}(A) \triangleq \int_A \frac{\mu_{0|n}}{\mathrm{d}\varsigma}(x_0)\varsigma(\mathrm{d}x_0)$$

with, for $x_0 \in \mathsf{X}$ and $y_0 \in \mathsf{Y}$,

$$\frac{\mu_{0|n}}{\mathrm{d}\varsigma}(x_0) \triangleq \frac{W_0(x_0)L_0 \cdots L_{n-1}(x_0, \mathsf{X}^{n+1})}{\nu[g(\cdot, y_0)L_0 \cdots L_{n-1}(\mathsf{X}^{n+1})]}.$$

Lemma A.2. *Let $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{n+1})$. Then, for all $n \geq 0$ and $N \geq 1$,*

$$\phi_{\nu,0:n|n}^N f - \phi_{\nu,0:n|n} f = \sum_{k=0}^{n} \varphi_k^N(f),$$

where, for $k \geq 1$,

$$\varphi_k^N(f) \triangleq \frac{\sum_{i=1}^N \mathrm{d}\mu_{k|n}^N/\mathrm{d}\eta_k^N(\xi_{0:k}^{N,i})\Psi_{k,n}[f](\xi_{0:k}^{N,i})}{\sum_{j=1}^N \mathrm{d}\mu_{k|n}^N/\mathrm{d}\eta_k^N(\xi_{0:k}^{N,j})} - \mu_{k|n}^N \Psi_{k,n}[f], \tag{A.3}$$

$$\varphi_0^N(f) \triangleq \frac{\sum_{i=1}^N \mathrm{d}\mu_{0|n}/\mathrm{d}\varsigma(\xi_0^{N,i})\Psi_{0,n}[f](\xi_0^{N,i})}{\sum_{j=1}^N \mathrm{d}\mu_{0|n}/\mathrm{d}\varsigma(\xi_0^{N,j})} - \mu_{0|n}\Psi_{0,n}[f]$$

and the operators $\Psi_{k,n} : \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+1}) \to \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+1})$, $0 \leq k \leq n+1$, are, for fixed points $\widehat{x}_{0:k} \in \mathsf{X}^{k+1}$, defined by

$$\Psi_{k,n}[f](x_{0:k}) \triangleq \frac{L_k \cdots L_{n-1} f(x_{0:k})}{L_k \cdots L_{n-1}(x_{0:k}, \mathsf{X}^{n+1})} - \frac{L_k \cdots L_{n-1} f(\widehat{x}_{0:k})}{L_k \cdots L_{n-1}(\widehat{x}_{0:k}, \mathsf{X}^{n+1})}. \tag{A.4}$$

Proof. As a starting point, consider the decomposition

$$\begin{aligned}
&\phi_{\nu,0:n|n}^N f - \phi_{\nu,0:n|n} f \\
&\quad = \frac{\phi_{\nu,0}^N L_0 \cdots L_{n-1} f}{\phi_{\nu,0}^N L_0 \cdots L_{n-1}(\mathsf{X}^{n+1})} - \phi_{\nu,0:n|n} f \\
&\qquad + \sum_{k=1}^{n} \left[\frac{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1} f}{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1}(\mathsf{X}^{n+1})} - \frac{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1} f}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \right].
\end{aligned}$$

Using the definitions (A.1) and (A.2) of η_k^N and $\mu_{k|n}^N$, respectively, we may write, for $k \geq 1$,

$$\begin{aligned}
&\frac{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1} f}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \\
&\quad = \eta_k^N \left[\frac{W_k(\cdot) L_k \cdots L_{n-1} f(\cdot)}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \right] \\
&\quad = \eta_k^N \left[\frac{W_k(\cdot) L_k \cdots L_{n-1}(\cdot, \mathsf{X}^{n+1})}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \left\{ \Psi_{k,n}[f](\cdot) + \frac{L_k \cdots L_{n-1} f(\widehat{x}_{0:k})}{L_k \cdots L_{n-1}(\widehat{x}_{0:k}, \mathsf{X}^{n+1})} \right\} \right] \\
&\quad = \mu_{k|n}^N \left[\Psi_{k,n}[f](\cdot) + \frac{L_k \cdots L_{n-1} f(\widehat{x}_{0:k})}{L_k \cdots L_{n-1}(\widehat{x}_{0:k}, \mathsf{X}^{n+1})} \right] \\
&\quad = \mu_{k|n}^N \Psi_{k,n}[f] + \frac{L_k \cdots L_{n-1} f(\widehat{x}_{0:k})}{L_k \cdots L_{n-1}(\widehat{x}_{0:k}, \mathsf{X}^{n+1})}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\frac{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1} f}{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1}(\mathsf{X}^{n+1})} \\
&\quad = \frac{\sum_{i=1}^N \mathrm{d}\mu_{k|n}^N / \mathrm{d}\eta_k^N(\xi_{0:k}^{N,i}) \Psi_{k,n}[f](\xi_{0:k}^{N,i})}{\sum_{j=1}^N \mathrm{d}\mu_{k|n}^N / \mathrm{d}\eta_k^N(\xi_{0:k}^{N,j})} + \frac{L_k \cdots L_{n-1} f(\widehat{x}_{0:k})}{L_k \cdots L_{n-1}(\widehat{x}_{0:k}, \mathsf{X}^{n+1})}
\end{aligned}$$

and, by combining the two latter identities, it follows from the definition (A.3) of $\varphi_k^N(f)$ that, for $k \geq 1$,

$$\varphi_k^N(f) = \frac{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1} f}{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1}(\mathsf{X}^{n+1})} - \frac{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1} f}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})}.$$

The identity

$$\varphi_0^N(f) = \frac{\phi_{\nu,0}^N L_0 \cdots L_{n-1} f}{\phi_{\nu,0}^N L_0 \cdots L_{n-1}(\mathsf{X}^{n+1})} - \phi_{\nu,0:n|n} f$$

can be verified in a similar manner. □

Note that, conditional on $\mathcal{F}_{k-1}^N$, the first term on the right-hand side of (A.3) is nothing but an importance sampling estimate of $\mu_{k|n}^N \Psi_{k,n}[f]$, based on N independent η_k^N-distributed variables.

Lemma A.3. *Assume* (A1) *and let, for* $n \geq 0$ *and* $0 \leq i \leq n$, $f_i \in \mathcal{C}_i(\mathsf{X}^{n+1})$. *Furthermore, let, for* $k \geq 0$, *the operator* $\Psi_{k,n}$ *be defined via* (A.4). *Then,*

$$\|\Psi_{k,n}[f_i]\|_{\mathsf{X}^{k+1},\infty} \leq 2\rho^{0\vee(i-k)}\|f_i\|_{\mathsf{X}^{n+1},\infty}.$$

Proof. For $k \geq i$, we bound $\Psi_{k,n}[f_i]$ from above by $2\|f_i\|_{\mathsf{X}^{n+1},\infty}$; however, for $k < i$, a geometrically decreasing bound of the function can be obtained by using the forgetting property of the conditional latent chain. Hence, by the Markov property of the posterior chain and using the definition of the forward kernels (see (2)),

$$\begin{aligned}\frac{L_k \cdots L_{n-1} f_i(x_{0:k})}{L_k \cdots L_{n-1}(x_{0:k}, \mathsf{X}^{n+1})} &= \mathbb{E}[f_i(X_{i:n})|X_{0:k} = x_{0:k}, \mathcal{G}_n] \\ &= \mathbb{E}[\mathbb{E}[f_i(X_{i:n})|X_i = x_i, \mathcal{G}_n]|X_k = x_k, \mathcal{G}_n] \\ &= \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\{x_k, \mathbb{E}[f_i(X_{i:n})|X_i = \cdot, \mathcal{G}_n]\}\end{aligned}$$

with $x_{0:k} \in \mathsf{X}^{k+1}$. Therefore, we may, for $k < i$, rewrite $\Psi_{k,n}[f_i](x_{0:k})$ as

$$\begin{aligned}&\Psi_{k,n}[f_i](x_{0:k}) \\ &\quad = \int_{\mathsf{X}} \{\mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}(x_k, \mathrm{d}x_i) - \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}(\hat{x}_k, \mathrm{d}x_i)\}\mathbb{E}[f_i(X_{i:n})|X_i = x_i, \mathcal{G}_n].\end{aligned}$$

Applying Theorem 3.1 to this difference yields

$$\begin{aligned}&|\Psi_{k,n}[f_i](x_{0:k})| \\ &\quad \leq 2\|\mathbb{E}[f_i(X_{i:n})|X_i = \cdot, \mathcal{G}_n]\|_{\mathsf{X},\infty}\|\mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}(x_k, \cdot) - \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}(\widehat{x}_k, \cdot)\|_{\mathrm{TV}} \\ &\quad \leq 2\|\mathbb{E}[f_i(X_{i:n})|X_i = \cdot, \mathcal{G}_n]\|_{\mathsf{X},\infty}\rho^{i-k} \leq 2\|f_i\|_{\mathsf{X}^{n+1},\infty}\rho^{i-k}.\end{aligned}$$

□

Lemma A.4. *Assume* (A1) *and let* $n \geq 0$. *Then, for all* $1 \leq k \leq n$, $x_{0:k} \in \mathsf{X}^{k+1}$, $y_k \in \mathsf{Y}$ *and* $N \geq 1$,

$$\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(x_{0:k}) \leq \frac{\|W_k\|_{\mathsf{X}^2,\infty}}{\mu g(y_k)(1-\rho)\sigma_-},$$

where η_k^N *and* $\mu_{k|n}^N$ *are defined in* (A.1) *and* (A.2), *respectively.*

Proof. First write, for $x_{0:k} \in \mathsf{X}^{k+1}$ and $y_{k+1} \in \mathsf{Y}$,

$$\begin{aligned}&L_k \cdots L_{n-1}(x_{0:k}, \mathsf{X}^{n+1}) \\ &\quad = \int_{\mathsf{X}} q(x_k, x_{k+1}) L_{k+1} \cdots L_{n-1}(x_{0:k+1}, \mathsf{X}^{n+1}) g(x_{k+1}, y_{k+1}) \mu(\mathrm{d}x_{k+1}) \\ &\quad \leq \sigma_+ \int_{\mathsf{X}} L_{k+1} \cdots L_{n-1}(x_{0:k+1}, \mathsf{X}^{n+1}) g(x_{k+1}, y_{k+1}) \mu(\mathrm{d}x_{k+1}). \qquad \text{(A.5)}\end{aligned}$$

Now, since the function $L_{k+1}\cdots L_{n-1}(\cdot,\mathsf{X}^{n+1})$ is constant in all but the last component of the argument,

$$\begin{aligned}
&L_{k-1}\cdots L_{n-1}(x_{0:k-1},\mathsf{X}^{n+1})\\
&\quad=\int_{\mathsf{X}} q(x_{k-1},x_k)g(x_k,y_k)\int_{\mathsf{X}} q(x_k,x_{k+1})L_{k+1}\cdots L_{n-1}(x_{0:k+1},\mathsf{X}^{n+1})\\
&\qquad\qquad\times g(x_{k+1},y_{k+1})\mu^{\otimes 2}(\mathrm{d}x_{k:k+1})\\
&\quad\geq \mu g(y_k)\sigma_-^2\int_{\mathsf{X}} L_{k+1}\cdots L_{n-1}(x_{0:k+1},\mathsf{X}^{n+1})g(x_{k+1},y_{k+1})\mu(\mathrm{d}x_{k+1}). \qquad \text{(A.6)}
\end{aligned}$$

Since the integrals in (A.5) and (A.6) are equal, the bound of the lemma follows. □

Proof of Proposition A.1. We start with (i). Since, conditional on $\mathcal{F}_n^N$, the random variables $f_i(\hat{\xi}_{0:n}^{N,j})$, $1\leq j\leq N$, are independent and identically distributed with expectations

$$\bar{\mathbb{E}}_\theta^N[f_i(\hat{\xi}_{0:n}^{N,j})|\mathcal{G}_n\vee\mathcal{F}_n^N]=\frac{1}{\Omega_n^N}\sum_{j=1}^N \omega_n^{N,j}f_i(\xi_{0:n}^{N,j}), \qquad \text{(A.7)}$$

applying the Marcinkiewicz–Zygmund inequality provides the bound

$$N^{p/2}\bar{\mathbb{E}}_\theta^N\left[\left|\frac{1}{N}\sum_{j=1}^N f_i(\hat{\xi}_{0:n}^{N,j})-\frac{1}{\Omega_n^N}\sum_{j=1}^N \omega_n^{N,j}f_i(\xi_{0:n}^{N,j})\right|^p\middle|\mathcal{G}_n\vee\mathcal{F}_n^N\right]\leq C_p\|f_i\|_{\mathsf{X}^{n+1},\infty}^p, \qquad \text{(A.8)}$$

where C_p is a universal constant depending only on p. Having control of this discrepancy, we focus instead on the L^p error associated with the weighted empirical measure $\phi_{\nu,0:n|n}^N$. We make use of the identity

$$a/b-c=(a/b)(1-b)+a-c$$

on each term of the decomposition provided by Lemma A.2. This, together with Minkowski's inequality, gives us the bound

$$\begin{aligned}
\|\varphi_k^N(f_i)\|_{p|\mathcal{G}_n\vee\mathcal{F}_k^N}\leq{}&\left\|\frac{1}{N}\sum_{j=1}^N\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\xi_{0:k}^{N,j})\Psi_{k,n}[f_i](\xi_{0:k}^{N,j})(k)-\mu_{k|n}^N\Psi_{k,n}[f_i]\right\|_{p|\mathcal{G}_n\vee\mathcal{F}_{k-1}^N}\\
&+\|\Psi_{k,n}[f_i]\|_{\mathsf{X}^{k+1},\infty}\left\|\frac{1}{N}\sum_{j=1}^N\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\xi_{0:k}^{N,j})-1\right\|_{p|\mathcal{G}_n\vee\mathcal{F}_{k-1}^N}. \qquad \text{(A.9)}
\end{aligned}$$

Applying the Marcinkiewicz–Zygmund inequality to the first term of this bound gives

$$N^{p/2}\bar{\mathbb{E}}^N\left[\left|\frac{1}{N}\sum_{j=1}^N\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\xi_{0:k}^{N,j})\Psi_{k,n}[f_i](\xi_{0:k}^{N,j})-\mu_{k|n}^N\Psi_{k,n}[f_i]\right|^p\middle|\mathcal{G}_n\vee\mathcal{F}_{k-1}^N\right]$$

$$\leq C_p \left\| \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k} \right\|^p_{\mathsf{X}^{k+1},\infty} \|\Psi_{k,n}[f_i]\|^p_{\mathsf{X}^{k+1},\infty} \tag{A.10}$$

and treating the second term in a similar manner yields

$$N^{p/2}\bar{\mathbb{E}}^N \left[\left| \frac{1}{N} \sum_{j=1}^N \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k}(\xi^{N,j}_{0:k}) - 1 \right|^p \middle| \mathcal{G}_n \vee \mathcal{F}^N_{k-1} \right] \leq C_p \left\| \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k} \right\|^p_{\mathsf{X}^{k+1},\infty}. \tag{A.11}$$

Thus, we obtain, by inserting these bounds into (A.9) and applying Lemmas A.3 and A.4,

$$\sqrt{N}\|\varphi^N_k(f_i)\|_{p|\mathcal{G}_n \vee \mathcal{F}^N_{k-1}} \leq 4C_p^{1/p}\rho^{0\vee(i-k)} \frac{\|W_k\|_{\mathsf{X}^2,\infty}\|f_i\|_{\mathsf{X}^{n+1},\infty}}{\mu g(y_k)(1-\rho)\sigma_-}. \tag{A.12}$$

For the first term of the decomposition provided by Lemma (A.2), we have, using the same decomposition technique as in (A.9) and repeating the arguments of Lemma A.4,

$$\begin{aligned} \sqrt{N}\|\varphi^N_0(f_i)\|_{p|\mathcal{G}_n} &\leq 2C_p^{1/p} \left\| \frac{\mu_{0|n}}{\mathrm{d}\varsigma} \right\|_{\mathsf{X},\infty} \|\Psi_{0:n}[f_i]\|_{\mathsf{X},\infty} \\ &\leq 4C_p^{1/p}\rho^i \frac{\|W_0\|_{\mathsf{X},\infty}\|f_i\|_{\mathsf{X}^{n+1},\infty}}{\nu g(y_0)(1-\rho)}. \end{aligned} \tag{A.13}$$

Now, (i) follows by a straightforward application of Minkowski's inequality together with (A.8), (A.12) and (A.13).

We turn to (ii). By means of the identity

$$a/b - c = (a/b)(1-b)^2 + (a-c)(1-b) + c(1-b) + a - c$$

applied to (A.3), we obtain the bound

$$\begin{aligned} &|\bar{\mathbb{E}}^N[\varphi^N_k(f_i)|\mathcal{G}_n \vee \mathcal{F}^N_{k-1}]| \\ &\quad\leq \|\Psi_{k,n}[f_i]\|_{\mathsf{X}^{k+1},\infty} \left\| \frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k}(\xi^{N,j}_{0:k}) - 1 \right\|^2_{2|\mathcal{G}_n \vee \mathcal{F}^N_{k-1}} \\ &\qquad + \left\| \frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k}(\xi^{N,j}_{0:k})\Psi_{k,n}[f_i](\xi^{N,j}_{0:k}) - \mu^N_{k|n}\Psi_{k,n}[f_i] \right\|_{2|\mathcal{G}_n \vee \mathcal{F}^N_{k-1}} \\ &\qquad \times \left\| \frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k}(\xi^{N,j}_{0:k}) - 1 \right\|_{2|\mathcal{G}_n \vee \mathcal{F}^N_{k-1}}. \end{aligned}$$

Thus, we get, by reusing (A.10) and (A.11),

$$|\bar{\mathbb{E}}^N[\varphi^N_k(f_i)|\mathcal{G}_n]| \leq \bar{\mathbb{E}}^N[|\bar{\mathbb{E}}^N_\theta[\varphi^N_k(f_i)|\mathcal{G}_n \vee \mathcal{F}^N_{k-1}]||\mathcal{G}_n]$$

$$\leq 4C_2\rho^{0\vee(i-k)}\frac{\|W_k\|^2_{\mathsf{X}^2,\infty}\|f_i\|_{\mathsf{X}^{n+1},\infty}}{N\{\mu g(y_k)\}^2(1-\rho)^2\sigma_-^2} \tag{A.14}$$

and treating the last term of the decomposition in a completely similar manner yields

$$|\bar{\mathbb{E}}^N[\varphi_0^N(f_i)|\mathcal{G}_n]| \leq 4C_2\rho^i\frac{\|W_0\|^2_{\mathsf{X},\infty}\|f_i\|_{\mathsf{X}^{n+1},\infty}}{N\{\nu g(y_0)\}^2(1-\rho)^2}. \tag{A.15}$$

Finally, from (A.7), we conclude that the multinomial selection mechanism does not introduce any additional bias and, consequently, (ii) follows from the triangle inequality, together with (A.14) and (A.15). □

Having established Proposition A.1, we are now ready to proceed with the proof of Theorem 3.3.

Proof of Theorem 3.3. Decomposing the difference in question yields the bound

$$\begin{aligned}\|\hat{\gamma}_n^{N,\Delta_n} - \gamma_n\|_{p|\mathcal{G}_n} \leq \sum_{k=0}^{n-1}\|\widehat{\phi}^N_{\nu,0:k(\Delta_n)|k(\Delta_n)}s_k - \phi_{\nu,0:k(\Delta_n)|k(\Delta_n)}s_k\|_{p|\mathcal{G}_n} \\ + \sum_{k=0}^{n-\Delta_n}|\phi_{\nu,0:k+\Delta_n|k+\Delta_n}s_k - \phi_{\nu,0:n|n}s_k|,\end{aligned} \tag{A.16}$$

where we have set $k(\Delta_n) = (k+\Delta_n)\wedge n$. By writing

$$\begin{aligned}&\mathbb{E}[s_k(X_k, X_{k+1})|X_{k+\Delta_n+1} = x_{k+\Delta_n+1}, \mathcal{G}_{k+\Delta_n}] \\ &\quad= \mathbb{E}[\mathbb{E}[s_k(X_k, X_{k+1})|X_{k+1} = x_{k+1}, \mathcal{G}_{k+\Delta_n}]|X_{k+\Delta_n+1} = x_{k+\Delta_n+1}, \mathcal{G}_{k+\Delta_n}] \\ &\quad= \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n}\cdots\mathrm{B}_{\nu,k+1|k+\Delta_n}(x_{k+\Delta_n+1}, \widehat{s}_{k|k+\Delta_n})\end{aligned}$$

with, for $x\in\mathsf{X}$,

$$\widehat{s}_{k|k+\Delta_n}(x) \triangleq \mathbb{E}[s_k(X_k, X_{k+1})|X_{k+1} = x, \mathcal{G}_{k+\Delta_n}],$$

we get that

$$\begin{aligned}&\phi_{\nu,0:k+\Delta_n|k+\Delta_n}s_k - \phi_{\nu,0:n|n}s_k \\ &\quad= \psi_{k+\Delta_n+1|k+\Delta_n}\mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n}\cdots\mathrm{B}_{\nu,k+1|k+\Delta_n}(\widehat{s}_{k|k+\Delta_n}) \\ &\qquad- \psi_{k+\Delta_n+1|n}\mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n}\cdots\mathrm{B}_{\nu,k+1|k+\Delta_n}(\widehat{s}_{k|k+\Delta_n}),\end{aligned}$$

where we have defined, for $\ell, m\geq 0$, $\psi_{\ell|m}\triangleq\mathbb{P}(X_\ell\in\cdot|\mathcal{G}_m)$. Hence, we obtain, using the exponential forgetting property (see Theorem 3.1) of the time-reversed conditional hidden

chain,

$$\begin{aligned}
&|\phi_{\nu,0:k+\Delta_n|k+\Delta_n} s_k - \phi_{\nu,0:n|n} s_k| \\
&\quad \le 2\|\widehat{s}_{k|k+\Delta_n}\|_{\mathsf{X},\infty} \|\psi_{k+\Delta_n+1|k+\Delta_n} \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}(\cdot) \\
&\qquad - \psi_{k+\Delta_n+1|n} \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}(\cdot)\|_{\mathrm{TV}} \\
&\quad \le 2\rho^{\Delta_n} \|s_k\|_{\mathsf{X}^2,\infty}.
\end{aligned} \tag{A.17}$$

Substituting (A.17) and the bound of Proposition A.1(i) into the decomposition (A.16) completes the proof of (i). The proof of part (ii) is entirely analogous and is omitted for brevity. □

A.2. Proof of Proposition 3.4

(A4) Let f_i be the function of Proposition A.1. For $p \ge 2$, $\ell \ge 1$, there exists a constant $\alpha_{p,\ell}^{(n)}(f_i) \in \mathbb{R}^+$ such that

$$\max\left\{\bar{\mathbb{E}}\left[\frac{\|W_k\|_{\mathsf{X},\infty}^p \|f_i\|_{\mathsf{X}^{n+1},\infty}^\ell}{\{\mu g(Y_k)\}^p}\right], \bar{\mathbb{E}}[\|f_i\|_{\mathsf{X}^{n+1},\infty}^\ell]; 0 \le k \le n\right\} \le \alpha_{p,\ell}^{(n)}(f_i).$$

Under assumption (A4), we have the following result.

Proposition A.5. *Assume* (A1) *and* (A2). *There then exist universal constants B_p and B, B_p depending only on p, such that the following holds true for all $N \ge 1$:*

(i) *if assumption* (A4) *is satisfied for $\ell = p \ge 2$, then*

$$\|\widehat{\phi}_{\nu,0:n|n}^N f_i - \phi_{\nu,0:n|n} f_i\|_p \le \frac{B_p[\alpha_{p,p}^{(n)}(f_i)]^{1/p}}{\sqrt{N}(1-\rho)}\left[\frac{1-\rho^i}{\sigma_-(1-\rho)} + \frac{n-i}{\sigma_-} + \frac{\rho^i}{(\mathrm{d}\nu/\mathrm{d}\mu)_-} + 1\right];$$

(ii) *if assumption* (A4) *is satisfied for $p = 2$, $\ell = 1$, then*

$$|\bar{\mathbb{E}}^N[\widehat{\phi}_{\nu,0:n|n}^N f_i - \phi_{\nu,0:n|n} f_i]| \le \frac{B\alpha_{2,1}^{(n)}(f_i)}{N(1-\rho)^2}\left[\frac{1-\rho^i}{\sigma_-^2(1-\rho)} + \frac{n-i}{\sigma_-^2} + \frac{\rho^i}{(\mathrm{d}\nu/\mathrm{d}\mu)_-^2}\right].$$

Proof. The proof of the first part is straightforward: combining Proposition A.1 and Minkowski's inequality provides the bound

$$\begin{aligned}
&\|\widehat{\phi}_{\nu,0:n|n}^N f_i - \phi_{\nu,0:n|n} f_i\|_p \\
&\quad = \bar{\mathbb{E}}^{1/p}[\|\widehat{\phi}_{\nu,0:n|n}^N f_i - \phi_{\nu,0:n|n} f_i\|_{p|\mathcal{G}_n}^p] \\
&\quad \le \frac{B_p}{\sqrt{N}(1-\rho)}\left\{\frac{1}{\sigma_-}\sum_{k=1}^n \bar{\mathbb{E}}^{1/p}\left[\frac{\|W_k\|_{\mathsf{X},\infty}^p \|f_i\|_{\mathsf{X}^{n+1},\infty}^p}{\{\mu g(Y_k)\}^p}\right]\rho^{0\vee(i-k)}\right.
\end{aligned}$$

$$+\frac{1}{(\mathrm{d}\nu/\mathrm{d}\mu)_-}\bar{\mathbb{E}}^{1/p}\left[\frac{\|W_0\|_{\mathsf{X},\infty}^p\|f_i\|_{\mathsf{X}^{n+1},\infty}^p}{\{\mu g(Y_0)\}^p}\right]+\bar{\mathbb{E}}^{1/p}[\|f_i\|_{\mathsf{X}^{n+1},\infty}^p]\Bigg\}.$$

We finish the proof by substituting the bounds of assumption (A4) into the expression above and summing up. The proof of the second part follows similarly. □

Proof of Proposition 3.4. The proof of the first part follows by applying Proposition A.5 and the bound (A.17) to the decomposition

$$\|\widehat{\gamma}_n^{N,\Delta_n}-\gamma_n\|_p\le\sum_{k=0}^{n-1}\|\widehat{\phi}_{\nu,0:k(\Delta_n)|k(\Delta_n)}^N s_k-\phi_{\nu,0:k(\Delta_n)|k(\Delta_n)}s_k\|_p$$
$$+\sum_{k=0}^{n-\Delta_n}\|\phi_{\nu,0:k+\Delta_n|k+\Delta_n}s_k-\phi_{\nu,0:n|n}s_k\|_p.$$

The second part is proved in a similar manner. □

Acknowledgements

This work was supported by a grant from the Swedish Foundation for Strategic Research, a French government overseas student grant and a grant from the French National Agency for Research (ANR-2005 ADAP'MC project). The authors are grateful to the anonymous referees who provided sensible comments on our results that improved the presentation of the paper.

References

[1] Andrieu, C. and Doucet, A. (2003). Online expectation–maximization type algorithms for parameter estimation in general state space models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* **6** VI69–VI72.

[2] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models.* New York: Springer. MR2159833

[3] Del Moral, P. (2004). *Feynman–Kac Formulae. Genealogical and Interacting Particle Systems with Applications.* New York: Springer. MR2044973

[4] Del Moral, P. and Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. *Séminaire de Probabilités XXXVII* (J. Azéma, M. Emery, M. Ledoux and M. Yor, eds.). *Lecture Notes in Math.* **1832** 415–446. Berlin: Springer. MR2053058

[5] Douc, R., Fort, G., Moulines, E. and Priouret, P. (2007). Forgetting of the initial distribution for hidden Markov models. Available at `http://arxiv.org/abs/math/0703836`.

[6] Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice.* New York: Springer. MR1847783

[7] Doucet, A., Godsill, J. and West, M. (2004). Monte Carlo smoothing for nonlinear time series. *J. Amer. Statist. Assoc.* **99** 156–168. MR2054295

[8] Fort, G. and Moulines, É. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* **31** 1220–1259. MR2001649

[9] Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *J. Finance* **42** 281–300.

[10] Jaquier, E., Polson, N.G. and Rossi, P.E. (1994). Bayesian analysis of stochastic volatility models (with discussion). *J. Bus. Econom. Statist.* **12** 371–417.

[11] Kitagawa, G. and Sato, S. (2001). Monte Carlo smoothing and self-organising state-space model. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.) 178–195. New York: Springer. MR1847792

[12] Kleptsyna, M.L. and Veretennikov, A.Y. (2007). On discrete time ergodic filters with wrong initial conditions. Technical report, Univ. Main and Univ. Leeds.

[13] Pitt, M.K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94** 590–599. MR1702328

[14] Ristic, B., Arulampalam, M. and Gordon, A. (2004). *Beyond the Kalman Filter: Particle Filters for Target Tracking.* Atrech House Radar Library.

[15] Wei, G.C.G. and Tanner, M.A. (1991). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

Received September 2006 and revised September 2007